AMRITA@FIRE-2014: Morpheme Extraction for Tamil using Machine Learning

Anand Kumar M, S Rajendran and K.P Soman

Centre for Excellence in Computational Engineering and Networking
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Coimbatore-641 112

m_anandkumar@cb.amrita.edu, kp_soman@amrita.edu

Abstract:

This article presents the working methodology of supervised Morpheme Extraction Task for Tamil language in Morpheme Extraction Task (MET) Task of FIRE-2014. In this attempt, Tamil Morphemes are extracted based on supervised machine learning algorithm, Support vector machines.

1 Introduction

The role of morphological analyzer is to identify root and morphemes of a word. Generally, rule based approaches are used for developing morphological analyzer system which are based on a set of rules and dictionary that contains root words and morphemes. Recently, machine learning approaches are found to be dominating the Natural Language Processing field. Machine learning is a branch of Artificial Intelligence (AI) concerned with the design of algorithms that learn from the examples. Here, supervised learning based morphological analyzer is implemented for Tamil verbs and nouns and a suffix based method is followed for proper nouns and pronouns.

2 Methodologies

Tamil language is morphologically rich and agglutinative. Such a morphologically rich language needs deep analysis at the word level to capture the meaning of the word from its morphemes and its categories. Each root is affixed with several morphemes to generate a word. In general, Tamil language is postpositionally inflected to the root word. Each root word can take more than ten thousand inflected word forms. Tamil language takes both lexical and inflectional morphology. Lexical morphology changes the meaning of the word and its class by adding the derivational and compounding morphemes to the root. Inflectional

morphology changes the form of the word and adds additional information to the word by adding the inflectional morphemes to the root.

The input to the morphological system is a POS tagged sentence. Initially, POS Tagged sentence is refined and divided according to the minimized POS tagset. Then morphologically analyzes the Noun (N) and Verb (V) forms using supervised machine learning algorithm. Pronouns (P) and Proper nouns (PN) are morphologically analyzed using suffix based method.

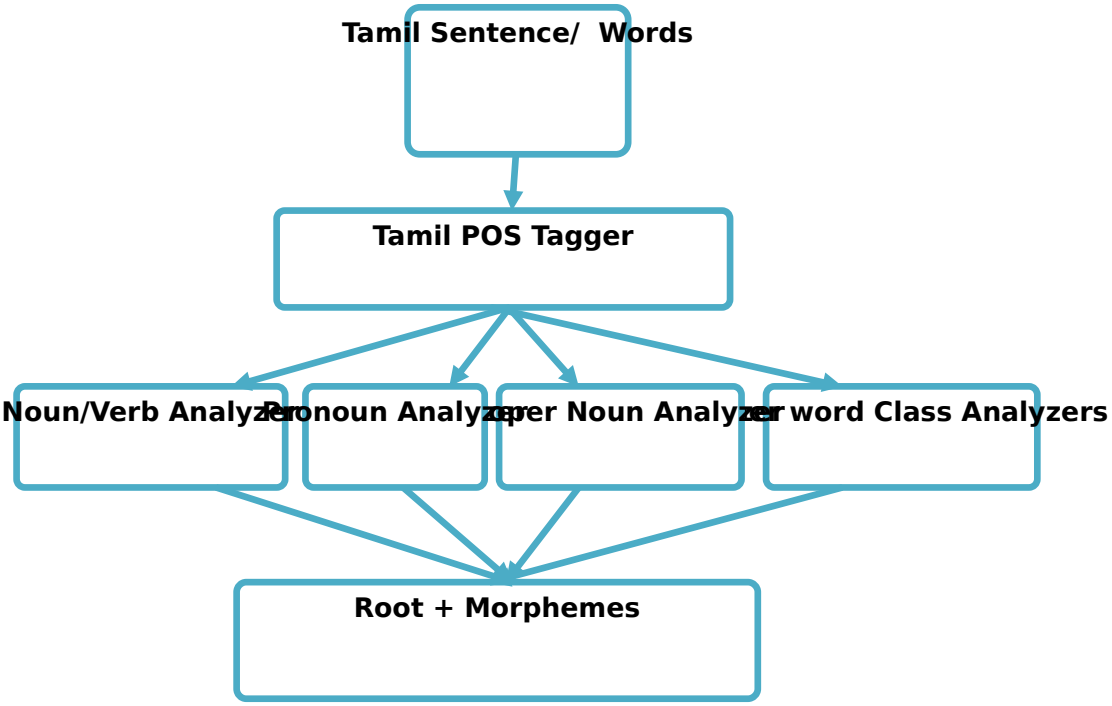


Figure 1: Framework of Tamil Morpheme Extraction

3 Machine learning based Morphological Analyzer for Nouns and Verbs

The sequence labeling is a significant generalization of the supervised classification problem. One can assign a single label to each input element in a sequence. The elements to be assigned are typically like parts of speech or syntactic chunk labels. Many tasks are formalized as sequence labeling problems in various fields such as natural language processing and bioinformatics. There are two types in sequence labeling approaches.

- Raw labeling.
- Joint segmentation and labeling.

In raw labeling, each element gets a single tag whereas in joint segmentation and labeling, whole segments get a single label. In the morphological analyzer, sequence is usually a word and a character is an element. Input word is denoted as ‘W’, and, root word and inflections are denoted by ‘R’ and ‘I’ respectively.

$$[W]_{Noun/Verb} = [R]_{Noun/Verb} + [I]_{Noun/Verb}$$

In turn, notation ‘I’ can be expressed as $i_1 + i_2 + \ldots + i_n$ where ‘n’ refers to the number of inflections or morphemes. Further ‘W’ is converted into a set of characters. Morphological analyzer accepts a sequence of characters as input and generates a sequence of characters as output. Let X be the finite set of input characters and Y be the finite set of output characters. If the input string is ‘x’, it is segmented as $x_1x_2....x_n$ where each $x_n \epsilon X$. Similarly, if y is an output string, it is segmented as $y_1y_2...y_n$ and yn ϵ Y where ‘n’ is the number of segments.

Inputs: $x = (x_1, x_2, x_3\ldots, x_n)$

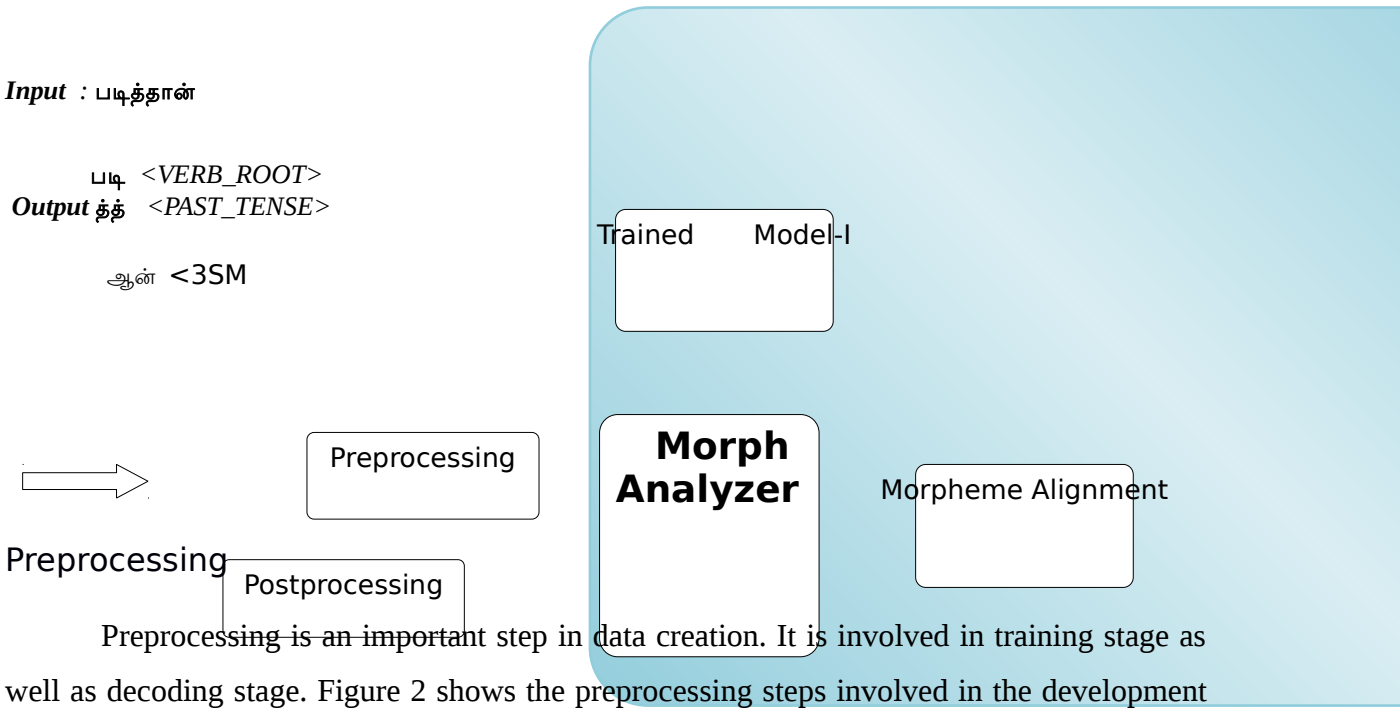
Labels: $y = (y_1, y_2, y_3\ldots, y_n)$

The main objective of sequence labeling approach is predicting y from the given ‘x’. In training data, the input sequence ‘x’ is mapped with output sequence ‘y’. Now the morphological analyzer problem is transformed into a sequence labeling problem. The information about the training data is explained in the following sub sections. Finally the morphological analysis is redefined as a classification task which is solved by using sequence labeling methodology.

3.1 Novel Data Modeling for Noun/Verb Morphological Analyzer

The goal of machine learning approach is to use the given examples and find out generalization and classification rules automatically. The data creation for the first phase of Noun/Verb Morphological analyzer system is done by preprocessing and mapping of characters. Data creation involved in the corpora development for morphological analyzer is classifying paradigms for verbs and nouns. The classification of Tamil verbs and nouns are based on tense markers and case markers respectively. Each paradigm will inflect with the same set of inflections. Paradigm provides information about all possible word forms of a root word in a particular word class. Tamil noun and verb paradigm classification is done

based on its case and tense markers respectively. Number of paradigms for each word class (noun/verb) is defined. The next step is to collect the list of root words for all paradigms .



Preprocessing is an important step in data creation. It is involved in training stage as well as decoding stage. Figure 2 shows the preprocessing steps involved in the development of corpora. Morphological corpus which is used for machine learning is developed by the following steps. Romanization, Segmentation and Alignment.

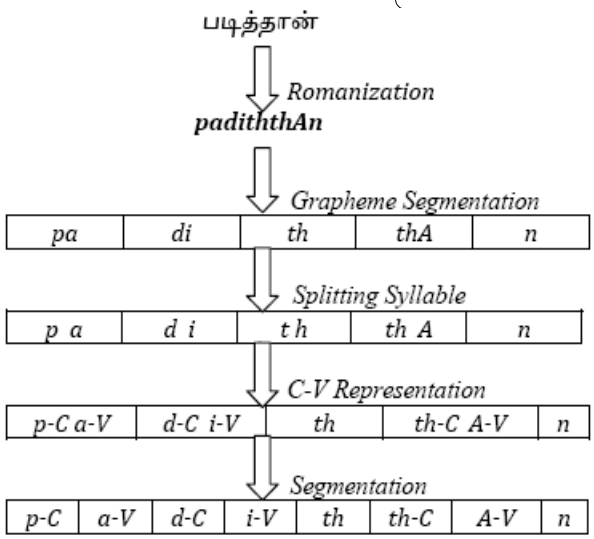


Figure 3 Preprocessing Steps

Table 1 Sample Data Format

I/P	*O/P*
p-C	**P**
a-V	**A**
d-C	**D**
i-V	**i***
th	**Th**
th-C	**th***
A-V	**A**
n	**n***

Segmentation of morpheme

In segmentation of morpheme process, words are segmented into morpheme according to their morpheme boundary. The input sequence is given to the trained Model-I. The trained model predicts each label to the input segments. This output sequence is aligned as a morpheme segments using alignment program.

Morpho-syntactic tagging

The segmented morpheme sequence is given to the trained Model-II. It predicts grammatical categories to the each segment (morphemes) in the sequence.

4 Suffix based Morphological Analyzer for Pronouns and Propernouns

The morphological analyzer for Proper Noun is developed by using the suffixes. Proper noun word form is taken as input for proper noun morphological analysis. Proper noun word form is taken from minimized POS Tagged sentence. It is identified from a POS tag <PN>. Initially proper noun word form is converted into roman form for easy computation. This Roman conversion is done by using simple key pair mapping of Roman and Tamil characters. This mapping program recognizes each Tamil character unit and replace with corresponding roman character. This Roman form is given to the proper noun analyzer system. System compares the word with the suffix which is predefined. First, it identifies the suffix and replaced with the corresponding information in the proper noun suffix data set. The suffix data set is created using various proper noun inflection and their end characters. For example, the word “sithamparam”(சிதம்பரம்) is end with ‘m’(ம்), and the other word “pANdisEri” (பாண்டிச்சேரி) is end with ‘ri‘(ரி) , the possible inflections of both words are given in table. Morphological changes are differing for the proper noun based on the end characters. So end characters are used in creating rules. From the various inflections of the word-form the suffix

is identified and the remaining part is stem. This suffix is mapped to the original morphological information. This algorithm replaces the encountered suffix with the morphological information in a suffix table.

5 Conclusion

This note explains the development of Morphological analyzer for Tamil language using Machine learning approach. Capturing the agglutinative structure of Tamil words by an automatic system is a challenging job. Generally, rule based approaches are used for building morphological analyzer. Tamil morphological analyzer for noun and verb is developed using the new and state of the art machine learning approach. Morphological analyzer problem is redefined as a classification problem. This approach is based on sequence labelling and training by kernel methods that captures the non linear relationships of the morphological features from training data samples in a better and simpler way. SVM based tool is used for training the system with the size of 6 lakh morphologically tagged verbs and nouns.

References

[1] John Goldsmith (2001), “Unsupervised Learning of the Morphology of a Natural Language”, Computational Linguistics, 27(2):153–198.

[2] Anandan. P, Ranjani Parthasarathy, Geetha T.V.2002. Morphological Analyzer for Tamil, ICON 2002, RCILTS-Tamil, Anna University, India.

[3] Daelemans Walter, G. Booij, Ch. Lehmann, and J. Mugdan (eds.)2004 , Morphology. A Handbook on Inflection and Word Formation, Berlin and New York: Walter De Gruyter, 1893-1900

[4] Dhanalakshmi V., Anand Kumar M., Rekha R.U., Arun Kumar C., Soman K.P., Rajendran S., "Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches," artcom, pp.433-435, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009

[5] Hal Daume , 2006. http://nlpers.blogsp-ot.com/2006/11/-getting-started-in-sequence-labeling.html

[6] S.Rajendran, Arulmozi, S., Ramesh Kumar, Viswanathan, S. 2001. “Computational morphology of verbal complex “.Language in india Volume 3 : 4 April 2003.

[7] Anand Kumar, M., et al. "A sequence labeling approach to morphological analyzer for Tamil language." *IJCSE) International Journal on Computer Science and Engineering* 2.06 (2010): 1944-195.